Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 306 (2003) 408–415
BBRC
www.elsevier.com/locate/ybbrc

A common periodic table of codons and amino acids

J.C. Biro,[a,b,*] B. Benyó,[c] C. Sansom,[d] Á. Szlávecz,[e] G. Fördös,[e] T. Micsik,[e] and Z. Benyó[e]

[a] *Karolinska Institute, Stockholm, Sweden*
[b] *Homulus Informatics, 88 Howard, #1205, San Francisco, CA 94 105-1649, USA*
[c] *Department of Informatics, Széchenyi István University, Egyetem tér 1., Györ 9026, Hungary*
[d] *Department of Crystallography, Birkbeck College, London WC1E 7HX, UK*
[e] *Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, Magyar tudósok körútja 2., Budapest 1117, Hungary*

Received 12 May 2003

Abstract

A periodic table of codons has been designed where the codons are in regular locations. The table has four fields (16 places in each) one with each of the four nucleotides (A, U, G, C) in the central codon position. Thus, AAA (lysine), UUU (phenylalanine), GGG (glycine), and CCC (proline) were placed into the corners of the fields as the main codons (and amino acids) of the fields. They were connected to each other by six axes. The resulting nucleic acid periodic table showed perfect axial symmetry for codons. The corresponding amino acid table also displaced periodicity regarding the biochemical properties (charge and hydropathy) of the 20 amino acids and the position of the stop signals. The table emphasizes the importance of the central nucleotide in the codons and predicts that purines control the charge while pyrimidines determine the polarity of the amino acids. This prediction was experimentally tested.

Keywords: Nucleic acid; Codon; Translation; Amino acid; Protein; Periodic table; Polarity; Charge; Hydrophobicity; Genetic code

The origin and development of the genetic code are not well understood. The 20 amino acids and the start and stop signals are coded redundantly by 64 codons. The amino acids (a.a.) may be classified according to basic physico-chemical properties, important ones being size, charge, and hydrophobicity. The four nucleic acids (n.a.) are either purines or pyrimidines and they form 64 different, simple patterns in the codons. We asked the question whether it is possible to find a relationship between the pattern of nucleic acids in the codons and the physico-chemical properties of the amino acids.

Two alternative hypotheses have frequently been posed to explain the origin of the genetic code. One hypothesis was championed by Woese [1], who argued that there was stereochemical matching—that is, affinity—between amino acids and certain triplet sequences. He therefore proposed that the genetic code developed in a way that was very closely connected to the development of the amino acid repertoire, and that this close biochemical connection is a fundamental of specific protein–nucleic acid interactions.

The other line was taken by Crick [2] who considered that the basis of the code might be a "frozen accident," with no underlying chemical rationale. He argued that the canonical genetic code evolved from a simpler primordial form that encoded fewer amino acids. The most influential form of this idea, "code coevolution" [3], proposed that the genetic code coevolved with the invention of biosynthetic pathways for new amino acids.

There are persuasive arguments for both theories [4,5]. The distribution of amino acid assignments found within the canonical genetic code is apparently nonrandom, however the connection between the code and the biochemical properties of an amino acid is not close either (or it is too sophisticated for our current understanding). Now, as we enter the post-genomic era, it is a more and more urgent challenge to understand the roles of macromolecular interactions. We believe that the

* Corresponding author.
E-mail addresses: jan.biro@kbh.ki.se (J.C. Biro), benyo@sze.hu (B. Benyó).

0006-291X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0006-291X(03)00974-4

code of the specific protein–protein and protein–nucleic acid interactions may be in the genetic code itself.

Methods and results

The influence of a single nucleotide on the codon was studied by translating homogeneous poly(A), poly(U), poly(G), poly(C) sequences around a single A, U, G, C residue (resulting single permutations of the four single-base codons). The almost universal Standard Genetic Code was used (transl_table = 1) [6]. Forty-eight codons produced in this way translate into 18 different amino acids and one stop signal. The remaining two amino acids—Asp (D) and Met (M) are only coded by codons that contain three different residues (Table 1). Phenylalanine (Phe), proline (Pro), lysine (Lys), and glycine (Gly) are unambiguously coded by homogeneous codons, while in other cases the presence and position of the single nucleotide result in more or less different translations. The position in the table of both hydrophobic and charged amino acids indicates that the pyrimidines (U and C) might control the property of hydropathy (hydrophobicity, hydrophilicity) while the purines (G and A) might determine the charge (positive, negative) of the amino acids. All codons coding charged amino acids contain at least one A or G.

It was necessary to construct a different periodic table to achieve a better separation of the physico-chemical properties of the amino acids and involve every codon, even those containing three different bases (Table 2). The four main codons UUU (Phe), CCC (Pro), AAA (Lys), and GGG (Gly) were placed into the corner of four fields each containing 16 positions. They were connected to each other by two horizontal, two vertical, and 2 diagonal axes. Codons, containing only the two kinds of residues of the connected main codons ($6 \times 6 = 36$ altogether), were placed along the connecting axes. The remaining 24 codons, including those that contain three different bases, were placed into the remaining places, where they fitted best. The table constructed by that way is periodical for the 64 possible codons and the contributing four nucleotides. It shows multiple axial symmetries (Table 3).

It is possible to construct three different periodic tables in this way, depending on which nucleotide in the codon (first, central or third) was chosen to organize the connections along the connecting axes. Using the second periodic table of codons that was organized around the second (middle) nucleotide in the codon produced a periodic table even for the corresponding amino acids, using the first or the third did not. Therefore we continued to work on this table only and we call it *A Common Periodic Table of Codons and Amino Acids.*

There are four main fields (or families of codons) in this table: xUx, xCx, xAx, and xGx. The allocation of the four main physico-chemical properties (hydrophobic, hydrophilic, positively charged, and negatively charged) and the stop signals is not random in this table, but shows periodicity.

- The xUx field contains exclusively hydrophobic amino acids.
- The xCx field is equally and symmetrically shared by hydrophilic and hydrophobic amino acids.
- The xUx plus xCx (pyrimidines) fields contain no charged amino acids but 20 of the 21 possible codons for hydrophobic amino acids, indicating that hydrophobicity is preferentially controlled by codons that contain mainly pyrimidine nucleotides.
- All charged amino acids and the stop signals are allocated into the xAx and xGx fields, indicating that the charge is preferentially determined by codons that contain mainly purine nucleotides.
- In the xAx and xGx fields the pattern of physico-chemical properties and stop signals is symmetrical, although this symmetry is not perfect.

Closely related amino acids might replace each other. Some replacements (Phe–Tyr, Glu–Asp, ...) are often found even in regions of high similarity and functional importance without any effect on the function of the protein. The most frequently occurring replacements are described in different "substitution frequency matrices" and "substitution tables." One such substitution table has been constructed by Turchin and Kohane [7]. The groups of amino acids in this table were found to occur together in columns of aligned sequences in both the BLOCKS [8] and HSSP [9,10] databases at a significant frequency, but they are separated from all other amino acids at a level of significance of $p < 0.01$. Some groups overlap each other in this substitution table because they contain amino acids that are common to more than one group. We grouped together the overlapping amino acids, resulting in three major groups and three non-substitutable amino acids (Table 4). The allocation of these amino acids is seen in Table 5. Amino acids belonging to the same group are located close to each other (neighbors) and form common, uninterrupted fields. The groups of substituting amino acids show a tendency to respect the borders of the four major fields in the periodic table.

The common periodic table of codons and amino acids indicates that hydrophobicity is controlled by pyrimidines, while charge is determined by purines. Three different types of proteins were chosen to test this prediction: (1) nuclear proteins, which are known to be charged sequences (the histones are positively while the acidic proteins are negatively charged under physiological conditions); (2) G-protein coupled receptors (GPCRs) which are known and expected to be hydrophobic because of their location in the cell membrane; and (3) protein ligands, each of which is known to bind specifically to its own GPCR. The example sequences in each group (10/group) were randomly selected from public sequence databases. The proportion of hydrophobic amino acids (L, V, I, F, M, P, A, W) and charged amino acids (K, R, E, D, H) in the peptide, and the proportion of the pyrimidine type nucleotides in the coding nucleic acid sequence are indicated in Fig. 1. There is a significant, positive correlation ($r = 0.876$, $p < 0.001$, $n = 30$) between the pyrimidine vs. purine nucleotide ratio in the coding nucleic acid sequences and the hydrophobic vs. charged amino acids ratio in the coded proteins (Fig. 2).

The nucleic acid and protein sequences of 10 randomly selected GPCR and their ligands (listed in Table 6) were examined to test the hypotheses that specifically interacting proteins (like a receptor and its ligand) might contain strings of amino acids that are coded by complementary codons. The coding sequences (CDS) of the 10 receptors were compared to the reversed and complemented (RC) sequences of their respective ligands before and after translation to all possible reading frames. The local sequence similarity searching methods BlastN, TblastX, BlastP [11], and BlastNP [12,13] were used for sequence comparisons. (BlastNP is an alternative method to TblastX and is defined as blastP applied on overlappingly translated nucleic acids. The method bypasses the frame-shift problematic.) The receptor sequences were not found to be similar to the reverse-complementary sequences of the ligands by any method (Table 6).

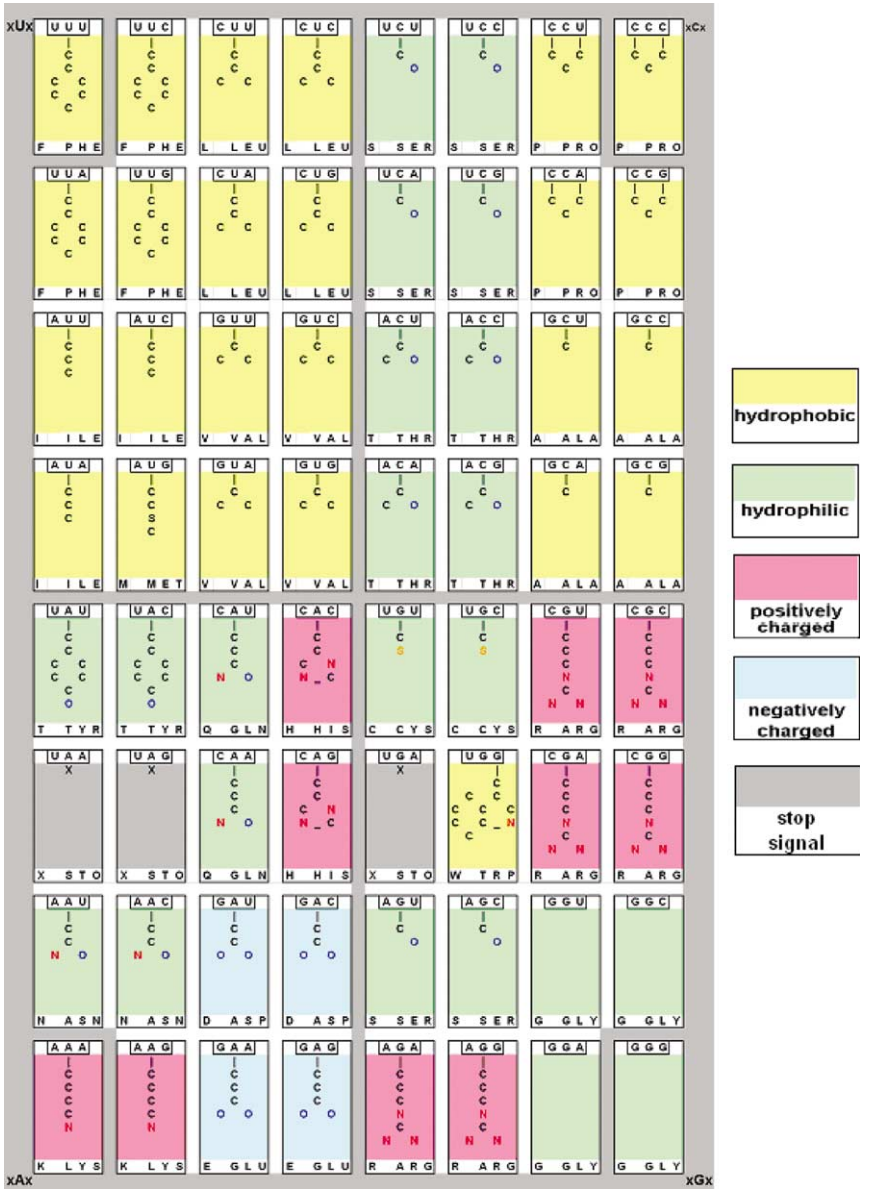
Discussion

The scientific effort to find a relationship between the nucleotide composition of codons and the biochemical properties of the coded amino acids is as old as the genetic code itself. The distribution of amino acid assignments found within the canonical genetic code is apparently non-random and there is a set of rules relating the nucleotide triplets to the amino acids. Some of these rules are obvious and have been known for a long time, for example, that all codons with a central U are cognate to amino acids with hydrophobic side chains, while codons with a central A are cognate to amino acids with polar side chains [14]. There is a weak

Table 1
The influence of a single nucleotide on the codon

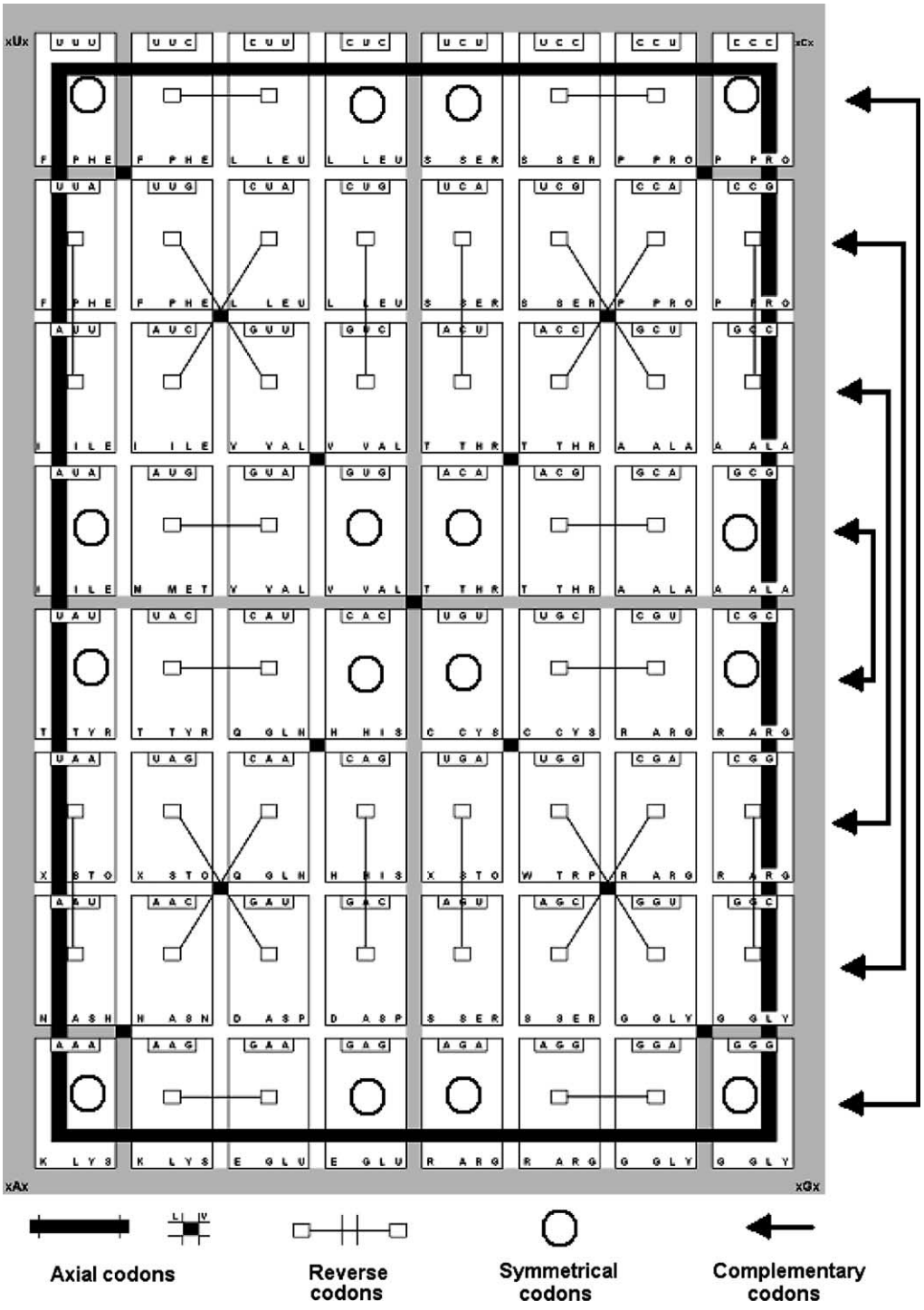
The frames of amino acid residues are rooted to the codons (boxes). The name of the amino acids is indicated by both one and three letters.

Table 2
Common periodic table of codons and amino acids



The frames of amino acid residues are rooted to the codons (boxes). The name of the amino acids is indicated by both one and three letters.

Table 3
Symmetries in the common periodic table of codons and amino acids



correlation between the ''symmetry'' of the codons and the hydropathy of the coded amino acids [15] as well as between the ''redundancy'' of the third nucleotide in the codon and the molecular weight of the coded amino acid [16,17]. Other patterns are more hidden and require the art of statistics to discover [18–22] and find ''the code within the codons.''

We have chosen a reductionist approach to start with and first tried to understand the effect of a single nucleotide change on the meaning of the codon. It was found that all but two amino acids and a stop signal could be coded by only two different nucleotides in the three-letter codon, and it was apparent that purines and pyrimidines played different and easily

Table 4
Substitution amino acid groups

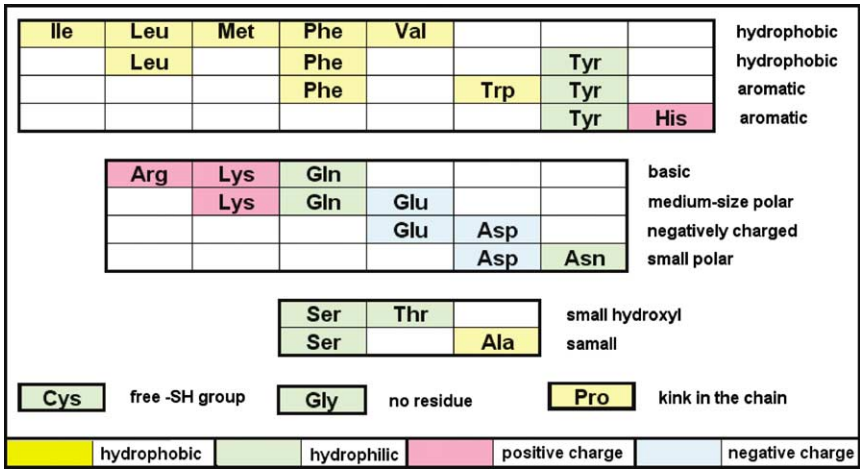
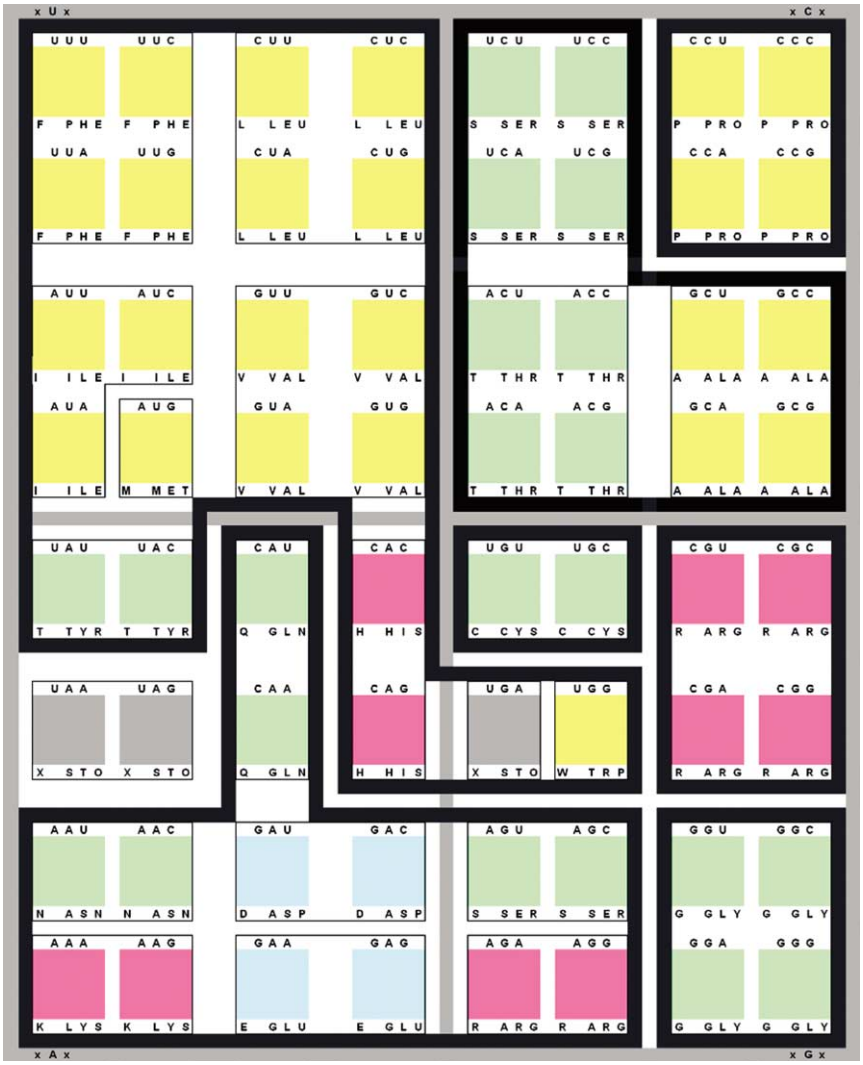


Table 5
Substitution amino acid groups in the codon periodic table



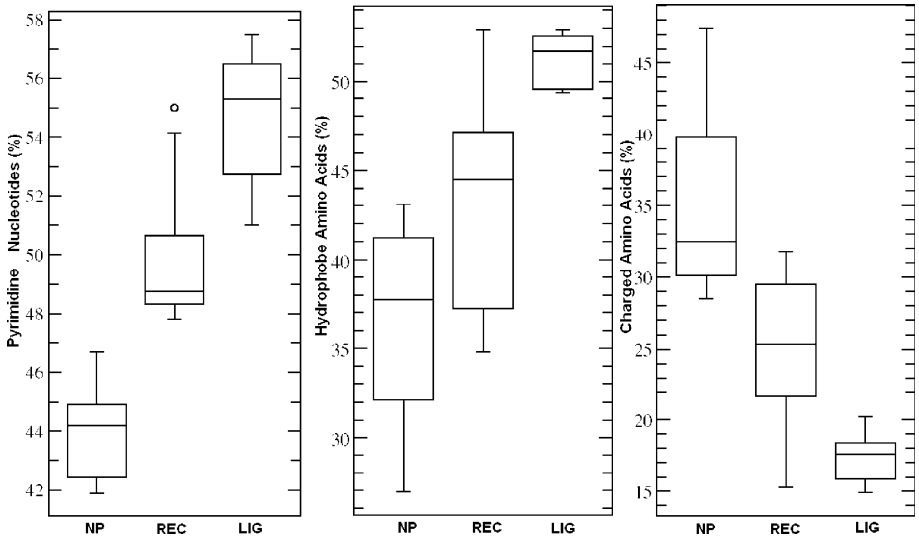


Fig. 1. Properties of three sequence types. The number of pyrimidine nucleotides (% of total) in the nucleic acid sequences and the number of hydrophobic and charged amino acids in the protein sequences (% of total) of selected nuclear proteins (NP), membrane receptors (REC), and receptor ligands (LIG); $n = 10$ in each group. The results are displayed using a Box Plot diagram.

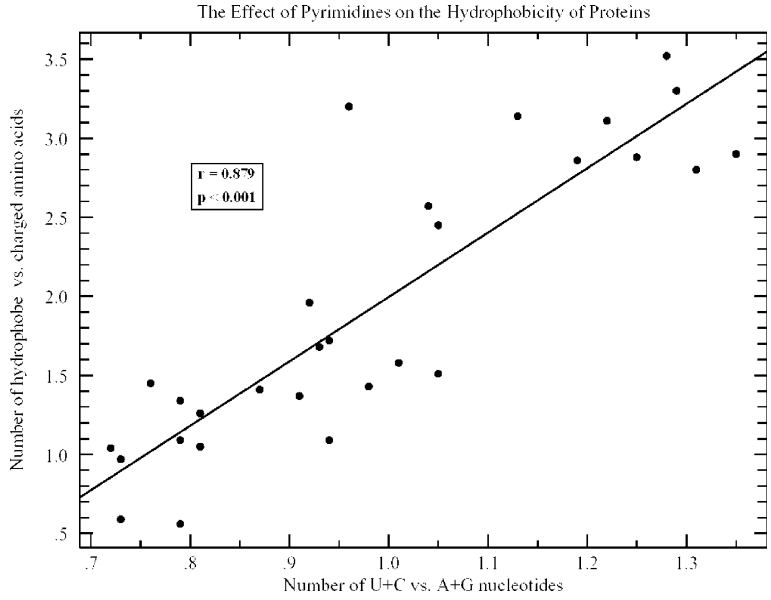


Fig. 2. The effect of pyrimidines in coding nucleic acids on the hydrophobicity of proteins. The correlation between the ratio of hydrophobic vs. charged residues in peptides and the ratio of pyrimidine (U + C) vs. purine (A + G) residues in the nucleic acids that code for those peptides. The sequences are the same as in Fig. 1; $n = 30$.

distinguishable roles in coding the properties of hydropathy and charge. This observation is well in agreement with previous studies [14]. The "single nucleotide in a homogeneous background" approach was also the cornerstone in our second study, where we aimed to "put in place" even those codons and amino acids that it was not possible to represent using the first approach. It was necessary to decide the topology (place in the triplet) of the "single nucleotides" that would define the four main fields of the codon periodic table. It was possible to construct three different codon periodic tables depending on whether the first, second or third nucleotide was chosen to organize the table. However, the physico-chemical properties of the amino acids showed the best periodicity when the second letter was chosen. The prominent role of the second nucleic acid in the codon in determining the properties of the coded amino acids has already been suggested by others [21,23,24], and our own results strongly support this view.

Table 6
Similarities between receptors and ligands

7TM-receptor (R)		Ligand (L)		L-frame/BLASTP-score				L/BLASTNP-score		L/BLASTN-score		
Name	Accession No.	Name	Accession No.	RC1	RC2	RC3	Pos. contr.	OTS-RC	Pos. contr.	N-D& N-RC	N-R& N-C	Pos. contr.
Angiotensin II receptor	M91464	Angiotensinogen	NM_000029	27	31	28	1865	103	3551	77	88	3626
Gastrin-releasing peptide receptor	M73481	Gastrin-releasing peptide	NM_002091	34	26	25	2008	91	3446	66	80	3422
Interleukin-8 receptor	M73969	Interleukin 8 (IL8)	NM_000584	25	26	34	1850	94	3081	81	95	2993
Endothelin B receptor	L06623	Endothelin 3	NM_000114	27	38	29	2340	97	3408	91	105	3408
Melanocortin-5 receptor	U08353	Propiomelanocortin	NM_000939	32	27	30	1682	101	2164	78	83	2081
Neuropeptide Y receptor	U36269	Neuropeptide Y	NM_000905	28	25	32	1956	93	2425	55	67	2379
Neurotensin receptor	Y10148	Neurotensin	NM_006183	28	36	31	2130	107	3169	61	89	3110
Somatostatin receptor	M81830	Somatostatin I	J00306	34	34	24	1913	78	2785	64	58	2678
Galanin receptor	AF080586	Galanin preproprotein	NP_057057	29	27	27	2038	105	2504	97	74	2401
Orexin receptor	AF041243	Orexin neuropeptide	NM_001524	26	30	27	2221	116	3237	87	84	3100

Readings of the sequences: D, direct; R, reverse; C, complementary; RC, reverse and complementary; Pos. contr., positive control, the receptor itself; N, nucleic acid; OTS, overlappingly translated sequence. Significant score: >10% of the Pos. contr.

It was already known that pyrimidines (especially U) in the central position of a codon specify hydrophobic amino acids. However, it was less well known that purines in the middle position might control the charge (positive as well as negative) of amino acids. The strong statistical correlation between the pyrimidine vs. purine ratio in the nucleic acids and the hydrophobic vs. charged amino acid ratio in the coded proteins suggests that this is truly the case. The clear separation of hydrophobic amino acids into the pyrimidine fields of our periodic table and charged amino acids into the purine fields is impressive.

Amino acids that often substitute for each other are neighbors in the periodic table. This further supports our observation that amino acids with similar properties group in the same field of the periodic table.

Some scientists have found that the properties of the amino acids showed a greater correlation with anticodonic than with codonic properties [25,26]. This is certainly not the case in our periodic table, because it is perfectly symmetrical for codons and anticodons.

The symmetric allocation of the codons–anticodons and some properties of the amino acids in the common periodic table invite us to examine an old but still exciting hypothesis: namely that there might also exist a form of amino acid complementarity in analogy to codon complementarity that might explain the nature of some specific protein–protein interactions [27–29]. We have not been able to confirm this theory before [30]. Now, we researched this possibility by comparing the nucleic acid or protein sequences of 10 membrane receptors to their specific ligands. The reverse-complementary sequences of the ligands (nucleic acid sequences and "protein sequences" in all possible reading frames) were not similar to the coding receptor nucleic acid sequences or the receptor proteins. Recent publications [31,32] indicating that synthetic protein translations of complementary nucleic acid sequences do specifically interact with each other remain an interesting observation.

It is challenging to interpret the possibility of constructing a common periodic table for codons and amino acids in the light of Woese's suggestion [1], and suppose that a strict stereospecific interaction between codons and the amino acids they code for has to exist and has always existed during evolution [33,34]. Indeed, there are some experimental evidences indicating that some amino acids specifically interact with their codons [35].

However a primitive set of codons (only one or two letters) might have "randomly" interacted with a primitive set of amino acids at the early phase of evolution, but then continued developing "logically" into the recent genetic code [36]. Even if the biosynthetic, coevolution theory seems to have serious weaknesses [37] Crick's hypothesis [2] is still not cracked and the "frozen accident" theory remains viable.

References

[1] C.R. Woese, in: The Genetic Code: The Molecular Basis for Gene Expression, Harper & Row, New York, 1967, pp. 156–160, Chapters 6–7.
[2] F.H.C. Crick, The origin of the genetic code, J. Mol. Biol. 38 (1968) 367–379.
[3] J.T. Wong, Evolution of the genetic code, Microbiology 5 (1988) 174–181.
[4] M. Yarus, RNA-ligand chemistry: a testable source for the genetic code, RNA 6 (2000) 475–487.
[5] T.A. Ronneberg, L.F. Landweber, S.J. Freeland, Testing a biosynthetic theory of the genetic code: fact or artifact? Proc. Natl. Acad. Sci. USA 97 (2000) 13690–13695.
[6] Available from: http://www3.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t.
[7] A. Turchin, I.S. Kohane, Gene homology resources on the World Wide Web, Physiol. Genomics 11 (2002) 165–177.
[8] Available from: http://www.blocks.fhcrc.org/.
[9] K.D. Pruitt, D.R. Maglott, RefSeq and LocusLink: NCBI gene-centered resources, Nucleic Acids Res. 29 (2001) 137–140.
[10] Available from: http://www.ncbi.nml.nih.gov/Database/index.html.
[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[12] J.C. Biro, Overlapping translation of nucleic acids for bioinformatics applications, in: The International Conference on Bioinformatics 2002 (InCB2002), Bangkok, Thailand, 2002.
[13] J.C. Biro, Overlapping translation of nucleic acids for bioinformatics applications, Med. Hypotheses 60 (2003) 654–659.
[14] M. Sjostrom, S. Wold, A multivariate study of the relationship between the genetic code and the physical–chemical properties of amino acids, J. Mol. Evol. 22 (1985) 272–277.
[15] C. Portelli, A.P. Portelli, D. Cigani, Symmetrical–asymmetrical codons and hydrophobic–hydrophilic amino acids, Physiologie 23 (1986) 139–143.
[16] M. Hasegawa, T. Miyata, On the antisymmetry of the amino acid code table, Orig. Life 10 (1980) 265–270.
[17] M. Di Giulio, Some aspects of the organization and evolution of the genetic code, J. Mol. Evol. 29 (1989) 191–201.
[18] J.D. Bashford, P.D. Jarvis, The genetic code as a periodic table: algebraic aspects, Biosystems 57 (2000) 147–161.
[19] I.Z. Siemion, Compositional frequencies of amino acids in the proteins and the genetic code, Biosystems 32 (1994) 163–170.
[20] G.I. Chipens, R.V. Rudzish, A model of intra- and intermolecular interactions of peptide chains, Ukr. Biokhim. Zh. 64 (1992) 20–28.
[21] S.J. Freeland, L.D. Hurst, The genetic code is one in a million, J. Mol. Evol. 47 (1998) 238–248.
[22] F.J. Taylor, D. Coates, The code within the codons, Biosystems 22 (1989) 177–187.
[23] V.A. Karasev, V.E. Stefanov, Topological nature of the genetic code, J. Theor. Biol. 209 (2001) 303–317.
[24] M.L. Chiusano, F. Alvarez-Valin, M. Di Giulio, G. D'Onofrio, G. Ammirato, G. Colonna, G. Bernardi, Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code, Gene 261 (2000) 63–69.
[25] J.R. Jungck, The genetic code as a periodic table, J. Mol. Evol. 11 (1978) 211–224.
[26] J. Konecny, M. Eckert, M. Schoniger, G.L. Hofacker, Neutral adaptation of the genetic code to double-strand coding, J. Mol. Evol. 36 (1993) 407–416.
[27] J. Biro, Comparative analysis of specificity in protein-protein interactions. Part II: The complementary coding of some proteins as the possible source of specificity in protein–protein interactions, Med. Hypotheses 7 (1981) 981–993.
[28] J.R. Heal, G.W. Roberts, J.G. Raynes, A. Bhakoo, A.D. Miller, Specific interactions between sense and complementary peptides: the basis for the proteomic code, Chembiochem 3 (2002) 136–151.
[29] A.E. Gabrielian, S. Heymann, Amino acid complementarity: testing of hypotheses, Biomed. Sci. 1 (1990) 311–313.
[30] U. Segersten, H. Nordgren, J. Biro, Frequent occurrence of short complementary sequences in nucleic acids, Biochem. Biophys. Res. Commun. 139 (1986) 94–101.
[31] I.Z. Siemion, R. Zbozien-Pacamaj, P. Stefanowicz, New hypothesis on amino acid complementarity and its evaluation on TGF-beta (2)-related peptides, J. Mol. Recog. 14 (2001) 1–12.
[32] J.R. Heal, G.W. Roberts, J.G. Raynes, A. Bhakoo, A.D. Miller, Specific interactions between sense and complementary peptides: the basis for the proteomic code, Chembiochem 3 (2002) 271.
[33] L.B. Hendry, E.D. Bransome Jr., M. Petersheim, Are there structural analogies between amino acids and nucleic acids? Orig. Life 11 (1981) 203–221.
[34] D. Grafstein, Stereochemical origins of the genetic code, J. Theor. Biol. 105 (1983) 157–174.
[35] M. Yarus, RNA-ligand chemistry: a testable source for the genetic code, RNA 6 (2000) 475–484.
[36] B.K. Davis, Evolution of the genetic code, Prog. Biophys. Mol. Biol. 72 (1999) 157–243.
[37] T.A. Rosenberg, L.F. Landweber, S.J. Freeland, Testing a biosynthetic theory of the genetic code: fact or artifact? Proc. Natl. Acad. Sci. USA 97 (2000) 13690–13695.